

微电子所等在数模混合存算一体芯片方面获进展

作者：writer 来源：中国科学院

本文原地址：<https://www.iikx.com/news/progress/26256.html>

本文仅供学习交流之用，版权归原作者所有，请勿用于商业用途！

当前，基于边缘智能计算设备运行的人工智能应用日趋复杂和高精度。为降低边缘设备运行的延迟和功耗，存算一体技术被应用在边缘设备端，通过减小数据搬运的开销最大化减少边缘设备上的延迟与功耗。而传统的存算一体宏仅支持使用整数型数据计算，难以支持日趋高精度、高复杂度以及片上训练的边缘端智能计算任务。仅使用单一模拟或数字方案的存算一体宏，难以在能量效率、面积效率和精度上取得最优化。如何有效结合模拟存算与数字存算模式优势，在总体上取得更高的能量效率和面积效率，并尽可能保证高精度，以及如何探索数模混合方案的设计空间，仍是存算一体宏领域亟需解决的问题。

中国科学院微电子研究所刘明院士团队等研发出基于外积运算的数模混合存算一体宏芯片，设计了数模混合浮点SRAM内存内计算方案，提出了模拟与数字存算宏的混合方法，结合了使用模拟存算方案进行高效阵列内位乘法和使用数字存算方案进行高效阵列外多位移位累加的优点，达到了整体上高能量效率与面积效率。研究通过残差式数模转换器架构，使数模转换器所需分辨率仅为输入位精度的对数，实现了高吞吐率和低开销。通过基于矩阵外积计算数学原理的浮点/定点存算块架构，矩阵-矩阵-向量计算可通过累加器元件完成。与之前的数字存算方案使用矩阵内积原理的大扇入、多级加法器树相比，可以降低运算的传输延迟，总的计算吞吐率更高。该架构支持细粒度的非结构激活稀疏性以进一步提升总体能效。该存算一体宏芯片在28nm CMOS工艺下流片，可支持BF16浮点精度运算以及INT8定点精度运算。BF16浮点矩阵-矩阵-向量计算峰值能效达到72.12TFLOP/W，INT8定点矩阵-矩阵-向量计算峰值能效达到111.17TFLOP/W。上述成果为采用数模混合方案的存算一体架构芯片提供了新思路。

近日，相关研究成果以A 28nm 72.12TFLOPS/W Hybrid-Domain Outer-Product Based Floating-Point SRAM Computing-in-Memory Macro with Logarithm Bit-Width Residual ADC为题，发表在2024国际固态电路会议（ISSCC 2024）上。该研究由微电子所和北京理工大学合作完成。研究工作得到国家重点研发计划、国家自然科学基金和中国科学院战略性先导科技专项等的支持。

研究团队单位：微电子研究所

更多 科学进展 请访问 <https://www.iikx.com/news/progress/>

本文版权归作者所有，请勿用于商业用途，[爱科学iikx.com](http://iikx.com)转发